Multimedia Lab @ ImageCLEF 2018 Lifelog Moment Retrieval Task

Mihai Dogariu and Bogdan Ionescu

Multimedia Lab, CAMPUS, University Politehnica of Bucharest, Romania
mdogariu@imag.pub.ro, bionescu@alpha.imag.pub.ro

Abstract. This paper describes the participation of the Multimedia Lab team at the ImageCLEF 2018 Lifelog Moment Retrieval Task. Our method makes use of visual information, text information and metadata. Our approach consists of the following steps: we reduce the number of images to analyze by eliminating the ones that are blurry or do not meet certain metadata criteria, extract relevant concepts with several Convolutional Neural Networks, perform K-means clustering on the Oriented Gradients and Color Histograms features and rerank the remaining images according to a relevance score computed between each image concept and the queried topic.

Keywords: Lifelog · CNN · Imagenet · Places365 · MSCOCO · Food101

1 Introduction

Recent technological advancements have resulted in the development of numerous wearable devices that can successfully help one track his own daily activity. Examples of such devices include wearable cameras, smart watches or fitness bracelets. Each of these provides information regarding its user's activity and combining the outputs of all such devices can result in a highly detailed description of the person's habits, schedule or actions. However, continuous acquisition of data can lead to cumbersome archives of information which, in term, can become too difficult to handle, up to the point where it becomes inefficient to try to use them. As part of ImageCLEF 2018 evaluation campaign [7], the Lifelog Tasks [4] aim to solve these problems.

This paper presents our participation in the Lifelog Moment Retrieval (LMR) task, in which participants have to retrieve a number of specific moments in a lifeloggers life, given a text query. Moments are defined as semantic events, or activities that happened throughout the day. For each query, a total of 50 images are expected to be extracted, both relevant and diverse, with the official metric being $F1@10$ measure.

The rest of the paper is organized as follows. In Section 2 we discuss related work from the literature, in Section 3 we present our proposed system, in Section 4 we discuss the results and in Section 5 we conclude the paper.

2 Related Work

In this section we briefly discuss the recent results obtained in similar competitions. The organizing team of the ImageCELF 2017 Lifelog Tasks [3] proposed a pipeline in which they perform a segmentation of the dataset based on time and concepts metadata. In parallel, they analyzed each query and extracted the relevant information that can be applied on the given metadata. After extracting only the images that fit the previous criteria they performed an additional filtering of images and remove those that contain large objects or are blurry. The last step involves a diversification of images through hierarchical clustering.

A similar technique was used by [12] in their submission at the same competition. In addition, they also used the image descriptors obtained by running each image through different Convultional Neural Networks (CNN), i.e. they extracted object and place feature vectors to which they added a human detection CNN. Each image was assigned a relevance score obtained by comparing the feature vector to a reference vector on a per topic basis. Their chosen clustering approach was K-means [11]. The same authors use a very similar system in [9], where they further add a temporal smoothing element. A somewhat different system was adopted in [17] where the authors combined a visual indexing method similar to the ones in [12, 9] with a location indexing method.

In our previous participation [5] we also applied a filtering procedure first based on the metadata and later on the similarity between the topic queries and the feature vector which consisted in detected concepts. This filtering was followed by a hierarchical clustering step. We learned that in order for this technique to work there has to be a strong correlation between the queries and the detected concepts. Also, enumeration of items that needed to be present in the image significantly improved the results.

This paper combines the benefits from [12] and [3] to which it adds two more feature vectors. Moreover, we explore the impact that supervised fine-tuning has on the final results and present the outcome of 5 different techniques.

3 Proposed Approach

Our approach involves the pipeline presented in Figure 1. Each of the processing steps is detailed in the following. The output of the system is a list of 50 images for each of the proposed 10 topics, which are both relevant and diverse with respect to the query.

3.1 Blur filtering

We first apply a blur filtering over the entire dataset. We compute a focus measure for each image by using the variance of the Laplacian kernel. If an image has a focus measure below an imposed threshold then it is discarded from further processing. Choosing the threshold requires several trials to see what works best for the dataset at hand. Imposing a low value on the threshold results in a

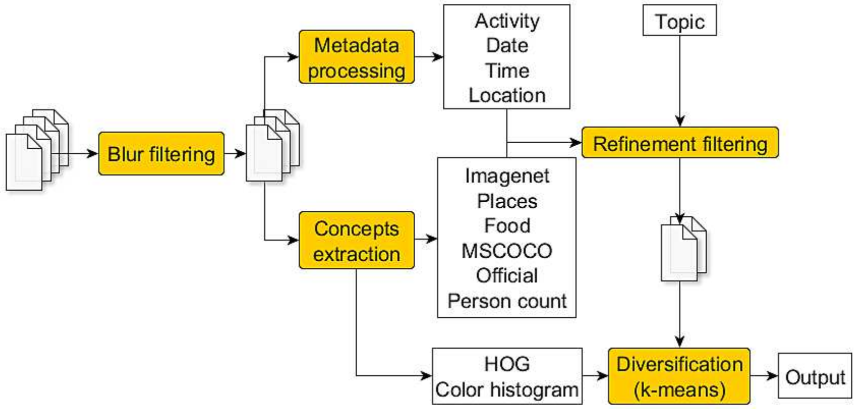

Fig. 1. Processing pipeline.

permissive filter, leading to a low number of discarded images, whereas a high threshold could wrongly discard images of acceptable quality. We found that a value of 60 for the threshold leads to satisfying results. We decided to allow the filter to be slightly permissive so that we do not reject true positives. In the end, from the total 80.5k images we discard 16.5k blurry images, leaving us with only 64k images to process. Another advantage of this technique is that it also filters out uninformative images that contain large homogeneous areas such as images where the camera was facing the ceiling or a wall.

3.2 Concepts extraction

In the second step of our algorithm we run each of the remaining 64k images through several classifiers and a detector. We use 3 image-level classifiers and one object detector, to which we also add the concept detector information provided by the organizers. All of these systems are implemented using CNNs as described below.

Imagenet classifier A common practice for detecting several concepts for an image is to run it through an image classifier trained on the popular Imagenet dataset [8]. This yields a 1000-D vector with values corresponding to the confidence level of associating the entire image with a certain concept. We use a ResNet50 [6] implementation trained on Imagenet.

However, there are 2 important aspects that need to be considered when implementing this technique. The first one is that the classifier is trained to predict a single concept for the entire image, whereas lifelog images contain numerous objects that might be of interest for the retrieval task. The second aspect is that out of the 1000 classes only a small part is relevant, with the vast majority of these concepts unlikely to be met in a person's daily routine. This leads to noisy classification, diminishing the usefulness of this classifier.

Places classifier The second classifier that we implement is meant to predict the place presented in the image. We use the VGG16 [14] network, trained on the Places365 dataset [18]. The dataset consists of approximately 1.8 million images from 365 scene categories. The network outputs a 365-D vector with one confidence value for each scene category. The places classifier performs well with respect to the lifelogging tasks, being trained to distinguish between most of the backgrounds present in the competition's dataset. This comes especially useful as most topics require the lifelogger to be present in a certain place at the time when the image has been captured.

Food classifier As some topics revolved around the lifelogger's eating and drinking habits we decided to also include a food classifier network. For this we use the InceptionV3 architecture [15] pre-trained on the Imagenet dataset and we fine-tune it on the Food101 dataset [1]. The result is a 101-D feature vector for each image. As the training dataset is composed of images where the labeled food takes up most of the image, when running our images through this classifier we extract 6 crops (upper left, upper right, center, lower left, lower middle, lower right) and their mirrored versions as well, which we pass through the network. Afterwards, we select the maximum activation for each food class from the 12 predictions and build the 101-D vector.

Object detector Additionally to the classifiers we also use a concept detector. This has the advantage that it locates more than one instance of the same object and each instance has its own attached confidence. Therefore, there will be no competition between detections when computing the final results. For this purpose we use a Faster R-CNN [13] implementation trained on the MSCOCO [10] dataset. Another advantage of this setup is that with object detection it also performs object counting. Therefore, we build two feature vectors for each image: one that retains the frequency of each detected object inside the image and one which sums up the confidences of all detected instances for each class inside the image. As the dataset also contains the class "person", we use its frequency to perform person counting. Also, many of the classes from the MSCOCO dataset can be found in daily scenarios, thus making it well-suited for the purpose of lifelog image retrieval.

Official concepts Apart from the previously mentioned systems there is one more feature extractor that we use, namely the one provided by the organizers. They released a set of results in which each image is described by a various number of concepts. The total number of possible classes is not known and their objective is also uncertain as they cover a broad range of concepts such as places, foods, actions, objects, adverbs etc. To cope with this we add each unique concept from the official feature results to a list that sums up 633 unique entries. In the end, we create a 633-D feature vector for each image, with non-zero entries only where the official concept detector triggered a detection. On this positions we retained the detector's confidence for the respective concept.

3.3 Metadata processing

Apart from the concept detector, the organizers also released a file containing a large variety of metadata about each minute from the logged data. These metadata encompass a bundle of information such as biometric data, timestamps, locations, activities, geographical coordinates, food logs and even the music that the lifelogger was listening to at certain times. We use only a part of this set of metadata. The rest of it can be used as well, but it did not fit our proposed system, therefore we only extract these data but did not process it any further. A summary of all the information that we process for each image can be seen in Table 1.

Table 1. Information used for individual images.

Type	Content	Dimension
Metadata	Activity	1-D
	Date	1-D
	Time (HH:MM:SS)	3-D
	Location	1-D
Concepts	Imagenet	1000-D
	Places	365-D
	Food	101-D
	MSCOCO objects	80-D
	MSCOCO person count	1-D
	Official concepts	633-D
Feature vectors	HOG descriptor	1536-D
	Color histogram	512-D

3.4 Refinement filtering

From previous experience we found that a key aspect of obtaining good results is to narrow down the set of images that are to be processed. This can be done by eliminating images that do not meet a certain set of minimum requirements. In this sense we implement two types of filtering: one based on the metadata and one based on the soft values of the concepts mentioned in Table 1 and explained below. We select a random topic out of the 10 test ones, to serve as an example and we will discuss it throughout the rest of the paper. The topic consists of the following:

Title: My Presentations
Description: Find the moments when I was giving a presentation to a large group of people.
Narrative: To be considered relevant, the moments must show more than 15 people in the audience. Such moments may be giving a public lecture or a lecture in the university.

Metadata filtering Our general approach is to manually interpret the entire topic text and extract meaningful constraints on the metadata associated with each image. Those entries that do not satisfy the given constraints are eliminated from the processing pipeline. We prefer looser restrictions such that we lower the chance of removing images relevant to the query in question. For the above given topic we impose the following:

- Activity: if the activity is any of the {'airplane', 'transport', 'walking'} then remove image;
- Location: if the location is anything different from {'Work', 'Dublin City University (DCU)'} then remove image;
- Time: if the hour is not in the interval 9-19 then remove image;
- Person count: if there are less than 10 persons detected then remove image. Two remarks are in order here. First, even if the person count is not part of the metadata we treat it as such because of its 1-D nature and discrete values. Second, the minimum threshold on the person count is lower than the query asks for because the MSCOCO object detector can have difficulties in detecting overlapped persons in an image.

Soft concepts filtering In a similar manner we tackle the filtering based on the soft outputs of the concept detector/classifiers. If a certain object/concept is detected with a higher probability than a preset threshold in an image then that image is removed from the processing queue. Again, this process involves manual selection of concepts that should not be present in the images. As it would be a tedious work to select an exhaustive set of concepts for each classifier, we only select the ones which are most likely to appear in the lifelog dataset and would be in contradiction with the queried text, therefore the selection can greatly differ from one query to another. For the query in the above example we select the following:

- Places: if the probability to detect any of the places from the set of words {'car_interior', 'living_room', 'kitchen'} is greater than the threshold then remove image;
- MSCOCO objects: if the probability to detect any of the objects from the set of words {'traffic light', 'cup'} is greater than the threshold then remove image;
- Official concepts: if the probability to detect any of the concepts from the set of words {'blurry', 'blur', 'null', 'Null','wall', 'ceiling', 'outdoor', 'outdoor_object'} is greater than the threshold then remove image;

We do not use the same technique for the Imagenet descriptor as it usually outputs low confidences and could thus have a great impact on the amount of images that would be removed. Also, the Food descriptor was not used for this topic as it is not relevant. Instead, its purpose is solely to classify food types for topics which implicitly ask for this.

We tried several values for the threshold and by visual inspection of the output we noticed that 0.3 offers a good trade-off between the probability of rejecting true positives and rejecting true negatives. Finding the best value for each concept detector and each topic requires many iterations, making this a costly process.

Relevance score After the blurred and irrelevant images have been filtered out we proceed into computing a relevance score for each image relative to the queried topic. In the same fashion as [12] we create a reference vector for each of the 5 concept detectors in Tabel 1 with higher values on the positions corresponding to concepts which are more likely to be found in relevant images and lower values on the other positions. The score associated to a certain concept detector is obtained by computing the dot product between the concept feature vector and its respective reference vector. The result is then weighted and added to the relevance score for each type of concept, as expressed in the equation below.

$$\begin{aligned}
score = & w_{imagenet} \times \sum_{i=1}^{1000} [concept_{imagenet}(i) * ref_{imagenet}(i)] + \\
& w_{places} \times \sum_{i=1}^{365} [concept_{places}(i) * ref_{places}(i)] + \\
& w_{food} \times \sum_{i=1}^{101} [concept_{food}(i) * ref_{food}(i)] + \\
& w_{mscoco} \times \sum_{i=1}^{80} [concept_{mscoco}(i) * ref_{mscoco}(i)] + \\
& w_{official} \times \sum_{i=1}^{633} [concept_{official}(i) * ref_{official}(i)],
\end{aligned} \tag{1}$$

with $concept_{<dataset>}(i)$ being the confidence associated with the i-th detected concept from a dataset for the respective image, $ref_{<dataset>}(i)$ being the reference vector at position i for the given dataset and $w_{<dataset>}$ being the weight given to the respective dataset. The weights for each dot product have been manually adjusted for each topic by trial and error. The values for the reference vectors have been either set manually or automatically, depending on the submitted run. We discuss this at length in Section 4.

3.5 Diversification

The submitted results are supposed to be both relevant and diverse. The relevance score should emphasize images that match the query description. For the diversity part we apply the K-means algorithm for all the images that are left after the filtering process. Each image is represented by the concatenation of two normalized vectors: a 1536-D vector representing the Histogram of Oriented Gradients (HOG) [2] feature vector and a 512-D vector representing the color histogram feature vector. This 2048-D vector should account for both shapes and colors inside images.

We run the K-means algorithm with either 5, 10, 25 or 50 clusters. For the final list of proposed images we select from each cluster the image with the highest relevance score in a round-robin manner.

4 Experimental Results

We have submitted one run during the competition and 4 other runs after the competition ended. The official metric of the competition was $F1@X$, which is computed as the harmonic mean between precision ($P@X$) and cluster recall ($CR@X$), with X representing the number of the top elements to be taken into consideration. In Table 2 we present the final $F1@X$ results that we have obtained for each run with best values in bold. Our last run is omitted when choosing the best results because it implied a highly supervised approach and would lead to an unfair comparison. In Figure 2 we present the $F1@X$ results for individual topics. Next, we provide a detailed description of each run.

Table 2. Official results for the submitted runs.

Run	F1@5	F1@10	F1@20	F1@30	F1@40	F1@50
Run 1	**0.235**	**0.216**	**0.224**	**0.218**	0.203	0.199
Run 2	0.154	0.169	0.215	0.21	**0.207**	0.199
Run 3	0.158	0.168	0.217	0.214	0.199	**0.206**
Run 4	0.129	0.166	0.184	0.184	0.178	0.188
Run 5	0.412	0.443	0.446	0.438	0.419	0.405

Run 1 This was the only run that we submitted during the competition and it follows the pipeline described in Section 3. We manually selected concepts from each training dataset that would be probable to appear in the images described by the queries. We set the reference vectors values to 1 on the positions corresponding to the selected concepts and to 0 elsewhere. This makes the dot product equivalent to an accumulation of confidences from a limited set of concepts for each image. The weights, $w_{imagenet}$, w_{places}, w_{food}, w_{mscoco} and $w_{official}$ have been adjusted independently for each topic. The official $F1@10$ value was 0.216 and this is the value that represents our position in the official standings.

Run 2 In addition to what was proposed for **Run 1** we also applied another filtering of the results, this time after the clusterization part. While going through the clusters in the round robbin manner we also checked that the newly added images are not too visually similar to the ones already added to the list. For this purpose each new proposal would be compared one-on-one with the already added proposals. The comparison was done with two metrics: mean squared error (MSE) and structural similarity index ($SSIM$). If for a pair of images $MSE < 2000$ and $SSIM > 0.5$ then they are considered to be too similar, the latter one is discarded and the round robin continues. We expected this technique to allow for more diversity in the proposed list of images and enhance the cluster recall. Instead, it turned out to eliminate a part of the correct predictions and lower the precision. The official $F1@10$ value was 0.169.

Run 3 For the 3rd run we proposed a different way of computing the reference vectors, the same technique that we used in [5]. Namely, instead of manually selecting the concepts that dictate whether an image is relevant or not from each dataset, we only selected the nouns that best describe the topic's description, obtaining a short set of key words, called "words_to_search". For the topic mentioned in Section 3.4 we have: *words_to_search={'presentation', 'group', 'people', 'audience', 'public', 'lecture', 'conference', 'university', 'classroom'}*. Starting from this set of words we computed the Wu-Palmer similarity measure [16] between each concept and all of the words from the "words_to_search" vector as described in the equation below.

$$ref_{dataset}(i) = \sum_{w \in words\_to\_search} d_{WUP}(concept_{dataset}(i), w), \tag{2}$$

where *dataset* is any of the 5 datasets used in the concept detectors (Imagenet, Places-365, Food-101, MSCOCO, Official), $d_{WUP}(concept_{dataset}, w)$ is the Wu-Palmer distance between one concept of the dataset and one word from the set of words to search for, "words_to_search" . This avoided the binary setting of the reference vector that was used in the previous runs but it lead to a decrease of the performance of the entire system. The official $F1@10$ value was 0.168.

Run 4 The 4th run was similar to **Run 3**, with the only difference being that all the weights $w_{imagenet}$, w_{places}, w_{food}, w_{mscoco} and $w_{official}$ were set to 1, rendering them neutral to the reference score computation. This allows the reference score to stabilize solely according to the similarity measure between the words from the topic description and the labels of the concept detectors. From Table 2 we can see that this only lowers the results, suggesting that tweaking the weights for each dot-product is a better approach. This run was our closest submission to a fully automatic system. The official $F1@10$ value was 0.166.

Run 5 Our last run was done with the same approach as **Run 1**, this time performing a fine-tuning of all system parameters for the topics that had bad results in the first run by trial and error. This approach leads to visibly better results. However, this is obtained after careful manual tuning, which makes the technique highly supervised and costly, as well, making it unfair to compare it with the previous runs, this being the reason why it is separated from the rest of the entries in Table 2. The official $F1@10$ value was 0.443.

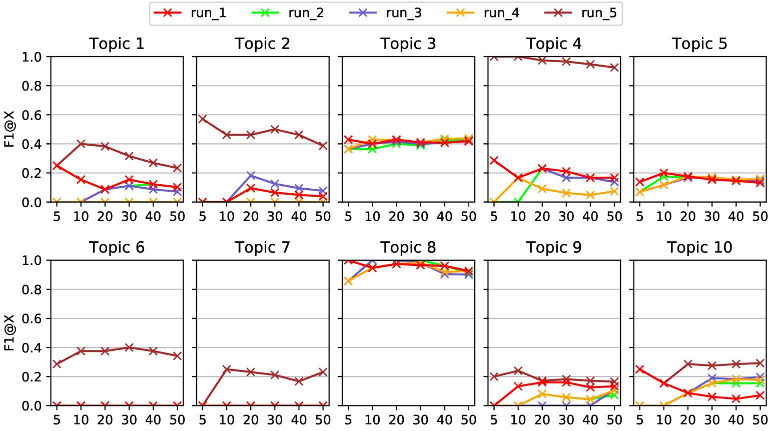

Fig. 2. Results for each topic from the test set.

4.1 Discussion

From the results that we presented in Figure 2 it can be seen that the $F1@X$ metric has high inter-topic variance. This does not come as a surprise since the topics approach different scenes, some of which are better represented in terms of number of images in the dataset or are better described in terms of the associated metadata. While some topics are easy to address (e.g. Topic 8: "Find the moments when I was with friends in Costa coffee." can be retrieved almost solely based on the location metadata) there are still topics for which retrieval is difficult (e.g. Topic 6: "Find the moments when I was assembling a piece of furniture.") mainly because of the difficulty of assigning distinctive concepts to their description. Except for the last run, it can be seen that all our approaches behave similarly for each individual topic, suggesting that there is no clear advantage in using one approach over the others. This is somewhat expected since they use the same data and almost the same degree of supervision. The only clear improvement can be seen when strong human input is involved.

The part of the entire system which had the greatest impact on the final outcome was the metadata filtering. We argue that this is because this type of information has been specifically implemented for lifelogging purposes and therefore have the strongest contribution in the end. This was also proven by our

5th run where we paid more attention to fine-tuning the processing parameters, such as metadata, weights and set of query words, rather than on introducing a new system.

The way the $F1@X$ metric changes with X is also worth mentioning. We noticed that is more beneficial to focus on the cluster recall than on the precision. This comes straightforward from the definition of the $F1@X$ metric in which $CR@X$ and $P@X$ have equal contributions. As the topics cover an average of 5-6 different clusters (as per the development dataset) it is usually more productive to retrieve images even from at least two different clusters rather than retrieve all the images from a single cluster. This happens because the cluster recall can only increase with X, whereas the precision usually drops for the same number of images. However, the cluster recall usually compensates for the precision.

We also notice that almost all of our approaches have the highest $F1@X$ value for $X = 20$ and they slightly decrease with the increase of X which was rather inconvenient since the official metric accounts for $X = 10$. However, we have reported quite similar results for $X = 10$ and $X = 20$.

5 Conclusions

In this paper we presented our approach for the LMR competition at the ImageCLEF Lifelog task. We have adopted a general framework that processes visual, text and meta information about images. We have extracted 5 concept vectors, 2 feature vectors and more than 10 metadata fields for each image. All of the proposed variants rely on metadata filtering and try to link each key-word from the search topics to the concepts detector labels. A relevance score which takes the aforementioned link into consideration is then computed and K-means algorithm is used for clustering the results for the final proposals.

The LMR task still poses numerous difficulties such as processing a great deal of multimodal data, adapting several multimedia retrieval systems to this type of task and integrating all the results. The diversity in the search queries is also to be taken into account, sometimes being quite easy to process (see results of ‘Topic 8’) but sometimes proving that there still is work to be done to find a solution that satisfies this type of generality (see results of ‘Topic 7’). We found that manual fine-tuning of system parameters offers the best result, but this makes the system personalized for the given topics, lowering its scalability to other similar tasks.

As opposed to last year, we have implemented a significantly more complex system and the future challenge for us is to work towards a scalable system, not so much dependent on human input, to solve the LMR task. We believe that with the increasing interest in this type of competitions it is possible to achieve this perspective.

Acknowledgement

This work was supported by the Ministry of Innovation and Research, UEFISCDI, project SPIA-VA, agreement 2SOL/2017, grant PN-III-P2-2.1-SOL-2016-02-0002.

References

1. Bossard, L., Guillaumin, M., Van Gool, L.: Food-101 – mining discriminative components with random forests. In: European Conference on Computer Vision. pp. 446–461 (2014)
2. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 1 - Volume 01. pp. 886–893. CVPR '05 (2005)
3. Dang-Nguyen, D.T., Piras, L., Riegler, M., Boato, G., Zhou, L., Gurrin, C.: Overview of ImageCLEFlifelog 2017: Lifelog Retrieval and Summarization. In: CLEF2017 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org <http://ceur-ws.org>, Dublin, Ireland (September 11-14 2017)
4. Dang-Nguyen, D.T., Piras, L., Riegler, M., Zhou, L., Lux, M., Gurrin, C.: Overview of ImageCLEFlifelog 2018: Daily Living Understanding and Lifelog Moment Retrieval. In: CLEF2018 Working Notes. CEUR Workshop Proceedings, vol. 11018. CEUR-WS.org <http://ceur-ws.org>, Avignon, France (September 10-14 2018)
5. Dogariu, M., Ionescu, B.: A Textual Filtering of HOG-based Hierarchical Clustering of Lifelog Data. In: CLEF2017 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org <http://ceur-ws.org>, Dublin, Ireland (September 11-14 2017)
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778 (June 2016)
7. Ionescu, B., Müller, H., Villegas, M., de Herrera, A.G.S., Eickhoff, C., Andrearczyk, V., Cid, Y.D., Liauchuk, V., Kovalev, V., Hasan, S.A., Ling, Y., Farri, O., Liu, J., Lungren, M., Dang-Nguyen, D.T., Piras, L., Riegler, M., Zhou, L., Lux, M., Gurrin, C.: Overview of ImageCLEF 2018: Challenges, datasets and evaluation. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018), vol. 11018. LNCS Lecture Notes in Computer Science, Springer, Avignon, France (September 10-14 2018)
8. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems 25, pp. 1097–1105 (2012)
9. Lin, J., Molino, A., Xu, Q., Fang, F., Subbaraju, V., Lim, J.H.: VCI2R at the NTCIR-13 Lifelog-2 Lifelog Semantic Access Task. In: Proceedings of the 13th NTCIR Conference on Evaluation of Information Access Technologies. Tokyo, Japan (December 5-8 2017)
10. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollr, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European Conference on Computer Vision (ECCV). vol. 8693, pp. 740–755. Zürich (2014)
11. Lloyd, S.: Least squares quantization in pcm. IEEE Trans. Inf. Theor. **28**(2), 129–137 (Sep 2006)

12. Molino, A., Mandal, B., Lin, J., Lim, J.H., Subbaraju, V., Chandrasekhar, V.: VC-I2R@ImageCLEF2017: Ensemble of Deep Learned Features for Lifelog Video Summarization. In: CLEF2017 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org <http://ceur-ws.org>, Dublin, Ireland (September 11-14 2017)
13. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: Cortes, C., Lawrence, N.D., Lee, D.D., Sugiyama, M., Garnett, R. (eds.) Advances in Neural Information Processing Systems 28, pp. 91–99. Curran Associates, Inc. (2015)
14. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: International Conference on Learning Representations (ICLR) (2015)
15. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016. pp. 2818–2826 (2016)
16. Wu, Z., Palmer, M.: Verbs semantics and lexical selection. In: Proceedings of the 32Nd Annual Meeting on Association for Computational Linguistics. pp. 133–138. ACL '94 (1994)
17. Yamamoto, S., Nishimura, T., Akagi, Y., Takimoto, Y., Inoue, T., Toda, H.: PBG at the NTCIR-13 Lifelog-2 LAT, LSAT, and LEST Tasks. In: Proceedings of the 13th NTCIR Conference on Evaluation of Information Access Technologies. Tokyo, Japan (December 5-8 2017)
18. Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A.: Places: A 10 million image database for scene recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **40**(6), 1452–1464 (June 2018)